

Learning from history: Non-Markovian analyses of complex trajectories for extracting long-time behavior

Ernesto Suarez and Daniel Zuckerman

Department of Computational and Systems Biology, University of Pittsburgh

July 8, 2014

Abstract

A number of modern sampling methods probe long time behavior in complex biomolecules using a set of relatively short trajectory segments. Markov state models (MSMs) can be useful in analyzing such data sets, but in particularly complex landscapes, the available trajectory data may prove insufficient for constructing valid Markov models. Here, we explore the potential utility of history-dependent analyses applied to relatively poor decompositions of configuration space for which MSMs are inadequate. Our approaches build on previous work [Suarez et. al., JCTC 2014] showing that, with sufficient history information, unbiased equilibrium and non-equilibrium observables can be obtained even for arbitrary non-Markovian divisions of phase space. We explore a range of non-Markovian approximations using varying amounts of history information to model the finite length of trajectory segments, applying the analyses to toy models as well as several proteins previously studied by μsec – msec scale atomistic simulations [Lindorff-Larsen et. al., Science 2011].

1 Introduction

The extremely complex dynamics of biomolecules can be difficult to sample and understand using straightforward molecular dynamics simulations, motivating the popularity of Markov state models (MSMs) which have been successfully applied to a number of systems [1, 2, 3, 4, 5]. However, MSMs can require extensive sampling and the careful definition of many states, and even so, the resulting modeling may exhibit non-Markovian behavior [5].

The question we investigate here is whether non-Markov models and analyses may be useful in some cases. We have shown in recent work [6] that using a subset of history information in the analysis of highly non-Markovian states is sufficient to build a model that gives unbiased state probabilities and kinetic properties. It is enough to introduce in the rate-matrix formulation and for each macroscopic transition, the last state visited. In other words, with sufficient history

information, it is not necessary to seek an optimal partitioning of the space or optimal slow coordinates as order parameters.

The main goal of this work is to explore the potential utility of non-Markovian analyses of trajectories. We try to obtain good kinetic estimates not only for arbitrary non-Markovian decompositions of configuration space, but also when the history information is limited in the sense that it is not possible to unambiguously assign the last state visited by a trajectory. That is, although we analyze very long trajectories, we consider non-Markovian approximations using varying amounts of history information to model the situation of finite trajectory segments. The analyses are applied to toy models as well as to several proteins previously studied by μsec – msec scale atomistic simulations [7].

2 Theoretical formulation

We compare a number of different analyses to well-sampled “brute force” trajectories, which provide reliable reference results. The focus here is on “poor” decompositions of configuration space which lead to non-Markovian behavior by construction.

Reference calculation of observables from long trajectories

All the simulations presented here are single regular brute force (BF) trajectories. The populations of configuration-space regions or “bins” are obtained by the trivial estimator $\hat{p}_i = c_i/C$, that turns out to be the maximum likelihood estimator (MLE), where c_i is the number of times the system was in bin i , and C is the total number of configurations. The first passage times (MFPTs) are measured and averaged directly during the simulation by just following the evolution of the trajectories.

Markovian calculation of observables

A traditional Markov analysis of the trajectories is performed for reference; this would be the case where no history information is included in the matrix analysis. The key quantity of interest is the presumed history-independent rate k_{ij} between two bins, defined by the conditional probability

$$k_{ij} = P\{X_{t+\tau} = j | X_t = i\}, \quad (1)$$

where X_t is the random variable representing the state of the system at time t , and τ is the lag-time used for the Markov model. Each rate is estimated by the MLE estimator

$$\hat{k}_{ij} = c_{ij}/c_i. \quad (2)$$

Here we are not using any constraints in the estimation, such as the one used for symmetric matrices [4], since the rates are reasonably well sampled in our analyses as can be inferred

by comparing the matrix results with direct BF measurements. The MFPTs are computed analytically as shown in reference [6].

Fully-history non-Markovian calculation of observables

This method has been previously described in detail [6] and uses the full history to label the trajectories depending of which is the last state visited. Suppose we have only two macroscopic states of interest A and B, which generally may not cover the full phase space. Every segment of a trajectory can be given a label according to whether the system was last in state A (the label α) or B (label β).

The labeled matrix approach explicitly uses the decomposition of the equilibrium population into α and β component for each bin i :

$$p_i^{eq} = p_i^\alpha + p_i^\beta. \quad (3)$$

Then, with N bins, a set of $2N$ probabilities is required. Similarly, a $2N \times 2N$ rate matrix is used for the analysis, and Eq. 2 is transformed into

$$\hat{k}_{ij}^\mu = c_{ij}^\mu / c_i^\mu, \quad \mu = \alpha, \beta, \quad (4)$$

where μ can be either label α or β depending on the last state visited. The bin probabilities and the MFPTs can be solved analytically [6].

Limited history analysis: Second-order Markov approach

We are also interested in modeling the case where a large set of unbiased trajectory segments is generated, perhaps from distributed computing or from a replica exchange simulation. To this end, we consider non-Markov analyses that employ time-limited history. Thus, for this particular analysis, we assume it is *not* possible to label all the trajectories – in the case there is more than one – as α or β .

Under limitation of finite history, a first approach could be to increase the order of the Markov approximation. The rates in a second order Markov model have three indexes, since they will depend on the state – in this case, the bin – the system occupied at times t and $t - \tau$:

$$k_{ij|m} = P\{X_{t+\tau} = j | X_t = i, X_{t-\tau} = m\}. \quad (5)$$

Similarly, the rates $\hat{k}_{ij|m} = c_{ij|m} / c_{i|m}$ are estimates from the transition counts, taking into account m , the bin occupied at $t - \tau$.

Numerical estimates of observables are obtained by kinetic simulation of the rate matrix.

Limited history analysis: Partial-color approach

We know that if it is possible to assign a “color” (label as α or β) to all the trajectories, the non-Markovian formulation is an unbiased way to estimate observables [6]. The main point of the partial-color method is to take advantage of the labeled information if it is available. Below, we examine analyses based on an amount of history equal to 5% and 10% of the average MFPT (average over forward and reverse directions). That is, when examining a given time point of the trajectory for estimating a labeled rate, the α or β label is only assigned properly if the trajectory occupied either of states A or B (as opposed to the intermediate region) during the preceding 5 or 10% of the trajectory.

Construction of the rate matrix proceeds in two steps. The first step is to build and solve a $2N \times 2N$ rate matrix where the rates are color independent, $k_{ij}^\mu = k_{ij}$, since that provides the Markovian values of p_i^α and p_i^β as well as default rate values. Next, each point in the trajectory is examined. If there is sufficient history to assign an α or β label, that is done. If not, the default Markovian values are used, in effect: the label μ is assigned to probability p_i^μ . This procedure is used to build the non-Markovian (labeled) $2N \times 2N$ count matrix $\mathbf{C} = \{c_{ij}^\mu\}$, that is transformed into a rate matrix (Eq. 4) used for the estimation of the observables.

3 Model systems and simulation details

We studied long trajectories for and protein folding systems generated by Shaw and coworkers [7], as well as simple toy models.

3.1 Toy models

Monte Carlo (MC) simulations were performed on two different toy models. The first is a one-dimensional model represented in Fig. 1. The space is divided in 10 bins, most of them of width π . The energy function in $k_B T$ units is given by

$$E_{1D} = \begin{cases} \sin(x) + 2.5 \cos(4x) + 0.0008x^4 - 0.11(x - 0.5)^2 & \text{if } -14 < x < 14 \\ \infty & \text{otherwise} \end{cases} \quad (6)$$

Two states A and B are defined as shown in Fig. 1. The state A consists of the first three bins while state B is the last four bins.

About 10^6 MC iterations were done for this model where the trial move δx is chosen randomly in the interval $[-\pi/2, \pi/2]$ with uniform probability distribution. Notice that the average displacement in x is $\pi/4$, four times smaller than the bin size.

The second model is the two-dimensional system shown in Fig. 2. The states A and B are defined by two bins in the configurational space, and the total number of bins used was 16 (see Fig. 2 right), each bin with dimensions $1.5\pi \times 1.5\pi$. The energy function in $k_B T$ units is given this time by Eq. 7.

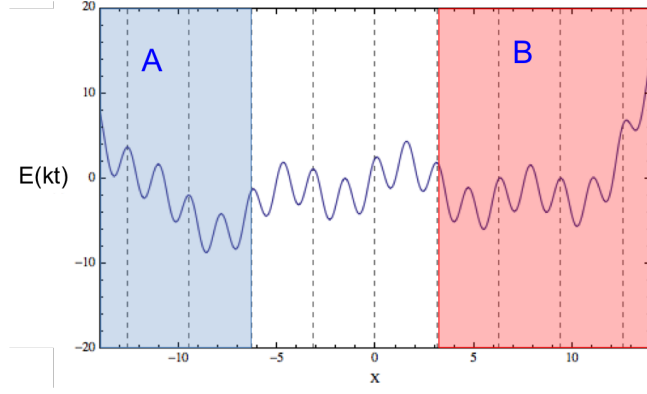


Figure 1: 1D toy model. The figure shows the definition of the states A and B and the partition of the space in bins is indicated with dashed lines.

$$E_{2D} = \begin{cases} -4.2 \sin(2x) \sin(y) + 0.02(y - (23 \sin(x/3) + x))^2 & \text{if } 0 < x, y < 6\pi \\ \infty & \text{otherwise} \end{cases} \quad (7)$$

The number of MC steps performed was about 2×10^6 , and the components of the trial moves δx and δy were selected independently and with uniform distribution in the interval $[-3\pi/8, 3\pi/8]$.

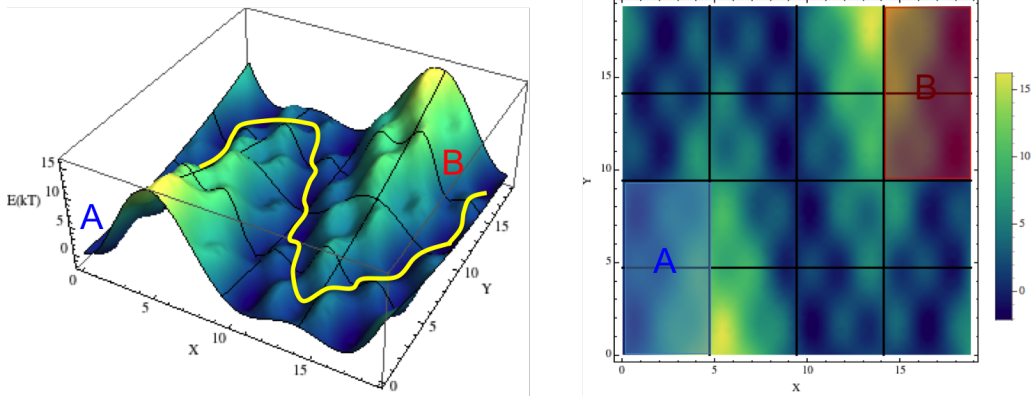


Figure 2: 2D toy model. Representation of the energy function and a possible path between A and B (left). Density plot indicating the partitioning of the space in bins and the definition of the states A and B in terms of bins (right).

A second partitioning of the space was also considered in order to create what we call pseudo one-dimensional model. In Fig. 2 there is a representation of what would be an “optimal path”, and the new binning tries to make bins almost perpendicular to that path, but of course, not in an optimal way. The goal here is to create a harder system to test the success of the methods described in Section 2.

3.2 Protein Models

Trajectories for five protein previously studied in explicit solvent by μsec – msec scale atomistic molecular dynamics simulations [7], were also analyzed using the approaches described in Sec-

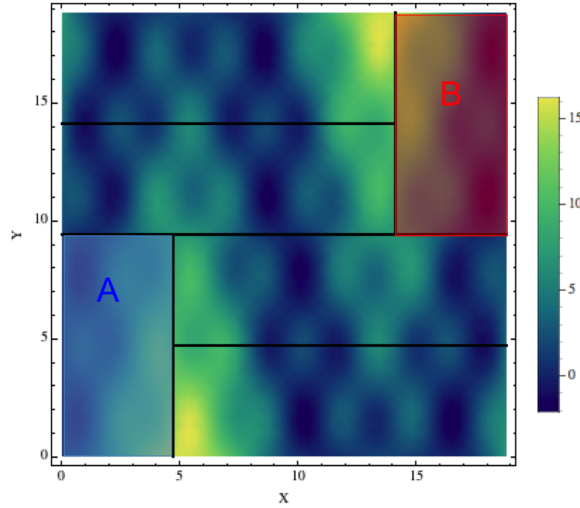


Figure 3: Pseudo 1D toy model

tion 2. We examined Chignolin, Trp cage, BBA, NTL9, and Villin. In the cases of BBA and NTL9, more than one trajectory was available, but we only analyzed the longest trajectory reported for each protein.

To create non-Markovian bins, the alpha-carbon RMSD (compared to the folded structure, excluding terminal residues) was used to define bins for the matrix analyses for all the proteins. All the trajectories were saved every $\tau=0.2\text{ns}$, and we used this same τ for the matrix analyses.

Table 1 shows, for each system, the definition of the states A (folded) and B (unfolded), as well as the number of bins used, the total simulation time considered for the analysis and the number of residues. Further details about the trajectories can be found in the original paper [7].

Table 1: Protein models used for Markovian and non-Markovian analyses. For each system, the table shows the number of residues, the total simulation time considered, the number of bins for the analyses and the states definitions.

Protein	Num. Residues	Time(μs)	Num. Bins	RMSD State A	RMSD State B
Chignolin	10	106	16	$< 2.0\text{\AA}$	$> 5.0\text{\AA}$
Trp-cage	20	208	9	$< 1.4\text{\AA}$	$> 6.0\text{\AA}$
BBA	28	225	6	$< 3.0\text{\AA}$	$> 7.0\text{\AA}$
NTL9	39	1100	7	$< 2.0\text{\AA}$	$> 5.0\text{\AA}$
Villin	35	125	10	$< 2.0\text{\AA}$	$> 6.0\text{\AA}$

4 Results

In this section we will show compare performance of different types of matrix analysis in terms of the MFPTs. The bin populations are also estimated from the matrices, but are not challenging to estimate from a Markovian formulation. The stationary distribution \mathbf{p} from a Markov model

obeys $K^T \mathbf{p} = \mathbf{p}$ and is equivalent to imposing the steady-state condition

$$\frac{d\mathbf{p}}{dt} = (K^T - I)\mathbf{p} = 0, \quad (8)$$

where K is the rate matrix and I , the identity matrix.

4.1 Toy models

The MFPTs were measured directly from the brute force trajectories, and also estimated from the matrix analyses described in Section 2. Since we are only interested in how the matrix analyses perform with respect to the direct BF measurement we divide all MFPT values by the corresponding BF direct measurement.

Fig. 4 and 5 show, respectively, the relative MFPTs from A to B and from B to A for the three toy systems: the one-dimensional toy model, the so-called pseudo one-dimensional model and the two-dimensional one.

In the plots legends “full history” means we are doing non-Markovian analysis, where we have enough history to label all of the trajectory according to the last of states A or B visited. On the other hand, “ $x\%$ MFPT” indicates partial color analysis where we are using only a history size of $x\%$ of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$. Finally, “2nd_Markov” is a second-order Markov approximation.

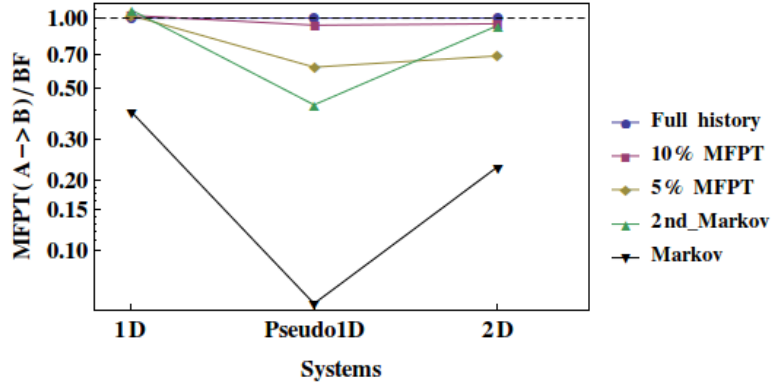


Figure 4: Relative MFPT from A to B, obtained from matrix analyses for the 1D, pseudo 1D, and the 2D models. The estimates are done using the full-history (non-Markovian), partial color analysis with 5% and 10% of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$, and second order Markov approximation.

4.2 Protein models

We applied the same analyses to the long-timescale trajectories generated by Shaw and coworkers [7]. As in the previous section, all the MFPT values plotted here are relative to the direct BF measurements. Fig. 6 and 7 show, respectively, the relative MFPTs from A to B and from B to A, obtain from for the five protein models considered.

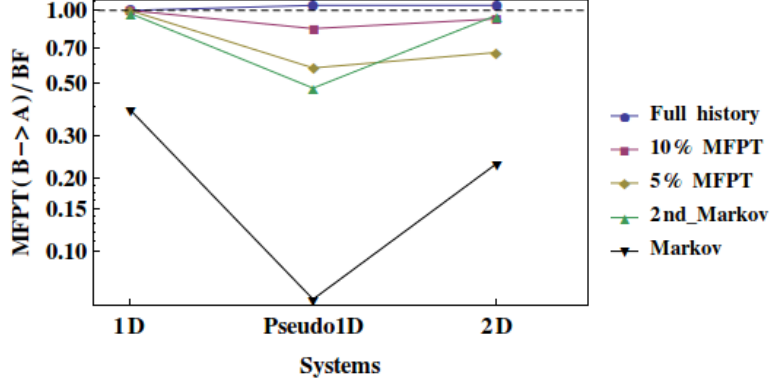


Figure 5: Relative MFPT from B to A, obtained from matrix analyses for the 1D, pseudo 1D, and the 2D models. The estimates are done using the full-history (non-Markovian), partial color analysis with 5% and 10% of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$, and second order Markov approximation.

Again, “full history” means we are doing non-Markovian analysis, where we have enough history to label all time-points in trajectory, “ $x\%$ MFPT” is a partial color analysis where only a history size of $x\%$ of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$ is taken into account, and “2nd_Markov” is a second order Markov approximation.

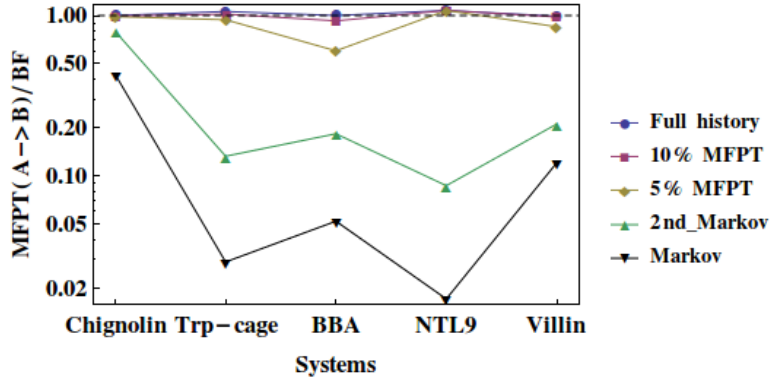


Figure 6: Relative MFPT from A to B, obtained from matrix analyses for protein models. The estimates are done using the full-history (non-Markovian), partial color analysis with 5% and 10% of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$, and second order Markov approximation.

5 Discussion and Conclusions

How can trajectory data be analyzed in the absence of a highly Markovian decomposition of phase space? This report presents an initial exploration of non-Markovian analyses applied to a wide range of systems, from toy models to proteins, in cases where phase-space decompositions were highly non-Markovian by construction. Perhaps the main conclusion is that a little history goes a long way: even a second-order Markov model with a short lag time provides reasonable estimates of first-passage times. An estimate can probably be considered reasonable if it is roughly within a factor of e from the true value, suggesting order $k_B T$ error in the effective barrier height — which is well within the accuracy limit of modern force fields [8, 9]. If further

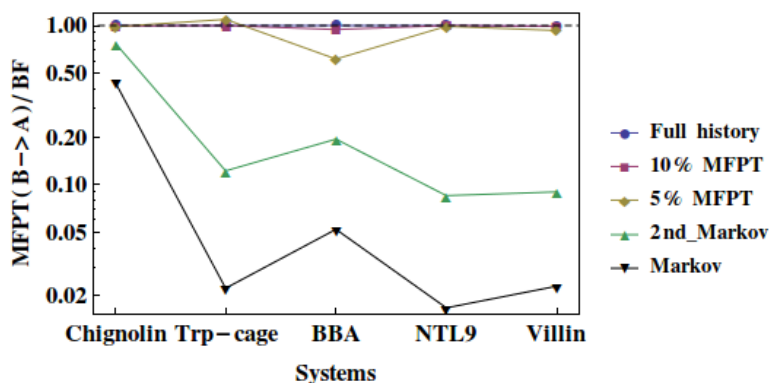


Figure 7: MFPTAB Relative MFPT from B to A, obtained from matrix analyses for protein models. The estimates are done using the full-history (non-Markovian), partial color analysis with 5% and 10% of $(\text{MFPT}(\text{AB}) + \text{MFPT}(\text{BA}))/2$, and second order Markov approximation

history information can be included, such as whether a trajectory previously visited one of the A or B macrostates of interest for rate estimation, better estimates can be obtained.

Although these results are encouraging, further studies should help to elucidate the general utility of these analyses. For example, how well do the analyses perform on less complete trajectories? At the same time, superior non-Markovian analyses may be within reach. It should not be difficult, for instance, to construct higher-order Markov models by employing progressively coarser decompositions of phase space further back into the history of a trajectory. Also, additional labeling information, including intermediate states/bins, may prove useful.

Acknowledgements

We thank Joshua Adelman for helpful discussions and the NSF for support (Grant Nos. MCB-0643456, MCB-1119091 and MCB-0845216).

References

- [1] Gregory R. Bowman, Daniel L. Ensign, and Vijay S. Pande. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J Chem Theory Comput*, 6(3):787–794, March 2010.
- [2] John D. Chodera, Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J Chem Phys*, 126(15):155101–17, 2007.
- [3] Nina Singhal Hinrichs and Vijay S Pande. Calculation of the distribution of eigenvalues and eigenvectors in markovian state models for molecular dynamics. *J Chem Phys*, 126(24):244101, Jun 2007.
- [4] Kyle A. Beauchamp, Gregory R. Bowman, Thomas J. Lane, Lutz Maibaum, Imran S.

- Haque, and Vijay S. Pande. Msmbuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J Chem Theory Comput*, 7(10):3412–3419, 2011.
- [5] Kyle A. Beauchamp, Robert McGibbon, Yu-Shan Lin, and Vijay S. Pande. Simple few-state models reveal hidden complexity in protein folding. *Proceedings of the National Academy of Sciences*, 109(44):17807–17813, 2012.
- [6] Ernesto Suarez, Steven Lettieri, Matthew C. Zwier, Carsen A. Stringer, Sundar Raman Subramanian, Lillian T. Chong, and Daniel M. Zuckerman. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J Chem Theory Comput*, 2014. In press, available online.
- [7] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, Oct 2011.
- [8] Michael R Shirts and Vijay S Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J Chem Phys*, 122(13):134508, Apr 2005.
- [9] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J Chem Phys*, 119:5740–5761, 2003.